Towards a Dataset of Activities for Action Recognition in Open Fields

Alexander Gabriel, Nicola Bellotto, Paul Baxter

Lincoln Centre for Autonomous System (L-CAS), University of Lincoln, UK

***Abstract* — In an agricultural context, having autonomous robots that can work side-by-side with human workers provide a range of productivity benefits. In order for this to be achieved safely and effectively, these autonomous robots require the ability to understand a range of human behaviors in order to facilitate task communication and coordination. The recognition of human actions is a key part of this, and is the focus of this paper. Available datasets for Action Recognition generally feature controlled lighting and framing while recording subjects from the front. They mostly reflect good recording conditions but fail to model the data a robot will have to work with in the field, such as varying distance and lighting conditions. In this work, we propose a set of recording conditions, gestures and behaviors that better reflect the environment an agricultural robot might find itself in and record a dataset with a range of sensors that demonstrate these conditions.**

I. Introduction

There are quite a number of datasets available that provide sensor readings of humans performing various activities. These usually come in the form of RGB videos, with ground truth in the form of action labels [1,2,3] or human skeletons (a set of joint positions organized as a graph) [2]. Today's datasets cover a wide variety of human actions, but mostly contain videos recorded by human camera operators under controlled lighting conditions. This results in videos where the subject is usually frame filling, conveniently oriented and illuminated well.

These conditions are not met in an agricultural setting where the camera operator is a robot, the camera can not be zoomed in on far-away targets or adjusted in direction, and the lighting conditions change with weather and the time of day. Additional problems can be caused by occlusions due to vegetation, infrastructure or machinery.

As a result of this mismatch, the researchers in [4] created a computer vision dataset with focus on the specific challenges for autonomous navigation in orchards like occlusions and poses uncommon in existing datasets.

Our research is carried out in the context of the RASberry project [5], which aims to develop autonomous fleets of robots for in-field transportation to aid and complement human fruit pickers.

In our setting, an agricultural robot has to cooperate with human field workers efficiently and comfortably. The workers pick berries into crates either in an open field or in a poly-tunnel. Once a crate is full, the robot will collect the crate and transport it to a destination outside the field for further processing.

This application requires basic communication between humans and robots. The robot has to learn where to go when, how far away from a picker it should stop and when it should leave again.

There are a number of interaction modes to chose from. Voice recognition, haptic interaction using buttons or touch screens, and gesture recognition either through remote observation or worn sensors have all been used in the past. We settled on remote gesture/behavior recognition as voice recognition is made infeasible by windy conditions and worn movement sensors as well as haptic interaction over distance rely on a wireless communication infrastructure that cannot be relied upon to be present in fruit fields.

Figure 1 shows how we are recording this dataset of action and behavior videos suitable to this task, i.e. in an open field at various distances and lighting conditions. We further extract skeletons using OpenPose [8,9] and investigate the influence of sensors and distances on extraction performance.

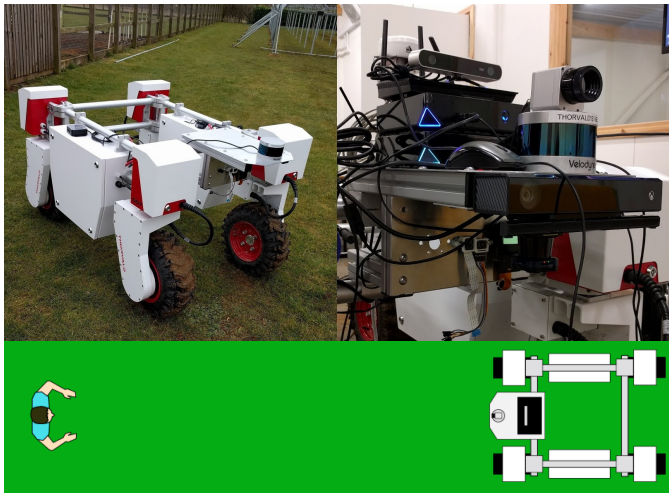

Figure 1: On the left: Our robot (SAGA Robotics Thorvald [6]) in front of our poly-tunnels. On the right: The sensor setup used in the recording. The figure at the bottom shows our experiment setup: An actor performing actions and behaviors at various distances to the robot.

In other work, OpenPose has been applied in a gait recognition task [10] and for human pose matching [11].

In Section II we will introduce the dataset in detail and motivate the design choices we made. In Section III we will give insight into the features of the dataset with special emphasis on the performance of different sensors at various distances, before concluding in Section IV.

II. Dataset Features

The choice of activities and decision to record at various distances are inspired by our application, the collection of fruit crates from human field workers and transportation of

said boxes to a cooling facility outside the field [5]. The dataset was recorded outside, on a piece of grassland, under varying lighting conditions (sunny, cloudy, morning to afternoon) and at distances ranging from 5m to 50m, at 5m intervals. Recording at different distances allows us to determine the performance of sensors and algorithms over the interaction range that the robot will face in action.

We recorded 10 actors, performing every activity once at each distance. Behaviors were performed from the front, back and side for a basic coverage of different directions.

The gestures were chosen for their relevance in basic communication between human and robot, the activities as a sample of interesting behavior displayed by human fruit pickers.

Thereafter each frame up to 25m distance was labeled with distance, actor ID, action and the direction the actor was facing. Labeling at further distances was hampered due to the actor being too small in the frame. The following list gives a short overview of dataset features:

- Distances: 5m - 50m at 5m intervals
- Actors: 10 actors, recorded individually
- Sensors: ZED stereo camera (RGB video and depth video), Optris thermal camera (thermal video), Velodyne VLP-16 (stream of 3D point clouds)
- Gestures: Waving, beckoning, indicating to stop, shooing, thumb up, thumb down, lower arm up, lower arm down, pointing
- Activities: Walking*, turning*, crouching down, standing up; with a crate in hand, *(marked classes also without crate)*

We chose a range of behaviors observable from human fruit pickers at work, and a set of gestures we deem helpful for basic communication over distances between 10 and 50 meters in the context of our application (i.e. directing a robot to collect and transport crates). The following two subsections will give a short overview of the gestures and activities.

A. Gestures

To direct the robot's attention to the worker in need of support, we selected a waving and a pointing gesture.

Waving: With the upper arm stretched out to the side, a rhythmic side to side motion of the lower arm.

Pointing: With the upper arm stretched out to the front, fist clenched except for the index finger which is also outstretched. We recorded this gesture at 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315° for basic directional coverage.

To facilitate comfortable and efficient loading of the robot, we want to direct it to a preferred stopping distance. For this we selected the beckoning, stop and shoo gestures.

Beckoning: With the arm partly stretched out to the robot and the palm facing the body, a sometimes circular fanning motion of the hand.

Indicating to stop: With the arm stretched out to the robot, the palm facing away from the body, fingers pointing up.

Shooing: With the arm partly stretched out to the robot and the palm facing the body, a fanning motion of the hand with emphasis (higher speed) on the motion away from the body.

For basic feedback purposes, we included a thumbs up/down gesture and a variant using the lower arm instead of the thumb, which should be easier to detect at further distances.

Thumbs up: With the arm partly stretched out to the robot, fist clenched and thumb sticking out, pointing up.

Thumbs down: With the arm partly stretched out to the robot, fist clenched and thumb sticking out, pointing down.

Lower arm up: With the upper arm stretched out to the side, the lower arm pointing straight up.

Lower arm down: With the upper arm stretched out to the side, the lower arm pointing straight down.

For sample stills of the recorded gestures, please refer to Figure 2 at the bottom of this page.

B. Activities

The most common activities in our domain are - besides the picking of berries - walking and turning around, crouching down, and standing up. Each of these activities occurs with free hands and while carrying a crate.

Being able to detect different behaviors allows the robot to learn activity models, specific to each individual worker, which allows it to predict the timing of future support requests.

Walking 5m with/without crate: Recorded from the front, the back, the right and left side.

Turning 90° with/without crate: Recorded from the front, the back and the right side.

Turning 180° with/without crate: Recorded from the front.

Crouching down with crate: Recorded from the front, the back and the right side.

Standing up with crate: Recorded from the front, the back and the right side.

Figure 2: A sample of the gestures we collected for the dataset. From left to right: wave, come, stop, shoo, thumb up, thumb down, lower arm up, lower arm down, pointing anti-clockwise at 45° intervals. The skeletons shown are 2D skeletons back-projected from 3D skeletons generated by the 'Lifting from the Deep algorithm' [7] run with OpenPose [8,9] 2D skeletons as input.

Table 1 shows the average duration for each action and behavior. The individual actions have a relatively short (<4s) duration and many of them like waving, shooing or the 'come' gesture consist of many, much shorter movements. A system running motion-based Action Recognition on this dataset will have to perform at a challenging framerate in order to capture these movements correctly.

TABLE I. AVERAGE DURATION PER ACTIVITY

Activity	*Average Duration [s]*	*Activity*	*Average Duration [s]*
Wave	3.73	Come	2.20
Shoo	2.22	Stop	2.25
Thumb up	1.71	Thumb down	1.90
Arm up	1.92	Arm down	2.09
Crate down away	1.83	Point 0°	1.92
Crate up away	1.29	Point 45°	1.91
Crate down side	1.21	Point 90°	2.00
Crate up side	1.30	Point 135°	1.82
Crate down toward	1.34	Point 180°	1.81
Crate up toward	1.11	Point 225°	1.88
Point 270°	1.99	Point 315°	1.63
Walk away (crate)	2.20	Walk away	3.20

III. DATASET CHARACTERIZATION

For the characterization of the dataset we combined the hand gesture classes (wave, come, stop, shoo, thumb up, thumb down) into a single class (hand gesture), as the skeleton models we use [7,8,9] do not support hand detection. Detection of individual fingers at longer distances is further complicated and ultimately prevented by sensor resolution.

The dataset was recorded outside which allows us to record at a wider range of distances and provides a natural variety in lighting conditions. The flat grassland, on which the dataset recording took place, is a well enough approximation for the flat ground we find in poly-tunnels, but does not feature enough occlusion of feet and lower legs or variations in ground level to model conditions in open fields.

Our data does not contain occlusions of the upper body except for self-occlusions from body parts/held items (crates). In this respect it is less challenging than the intended domain.

Recording outside allowed us to examine how larger distances affect skeleton extraction in our setup. Skeletons were extracted using OpenPose [8,9] from the RGB video as well as a color-coded version of the thermal camera feed. An example of the extracted skeletons is shown in Figure 3.

The confidence scores for skeleton extraction shown in Figure 4 are averages of the confidence scores produced by OpenPose for each skeleton. They are averaged over the duration of actions for different sensor sources individually.

The data shows significantly better skeleton extraction for action classes where the actor is facing the camera (arm down, arm up, wave, hand gestures, 'towards' gestures) compared to classes where the actor is facing to the side or away ('side' and 'away' gestures). This stems from self-occlusion of the further body side occurring in side views and self-occlusion of the arms by the torso when the actor is performing some action while facing away from the camera.

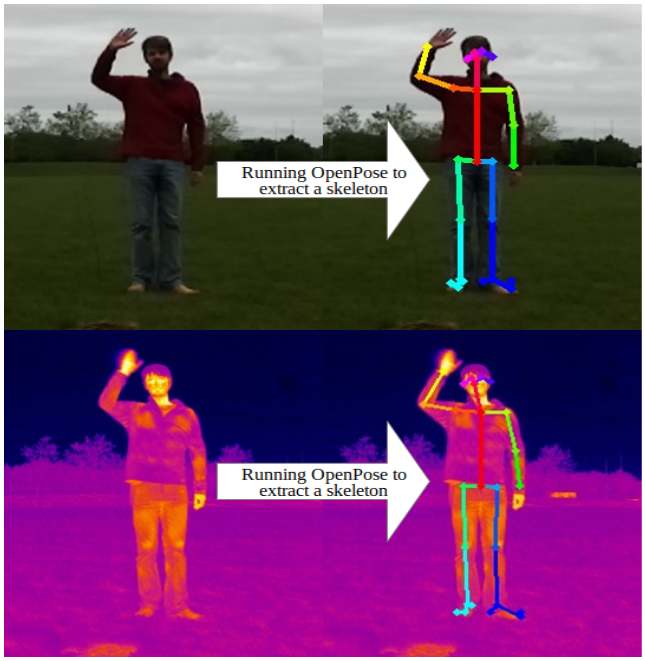

Figure 3: Results of running OpenPose on RGB video (top) and color-coded thermal video (bottom).

To note are also differences in scores for skeleton identification between the two sensors (see Figure 4), with the thermal source providing better skeleton identification for certain actions – a result that can be taken advantage of in the varying field conditions likely to be encountered.

Another interesting result are the generally higher scores for skeletons generated from RGB at close range combined with the lower scores for these skeletons at longe range. This validates our initial intuition that the wide-angle lens on the RGB-D camera would prove beneficial at short range but a disadvantage at longe range compared to the thermal camera. In general, skeleton extraction confidence tends to deteriorate at large distances for both sensors.

IV. CONCLUSION

Our experiments show a difference in skeleton extraction performance over the two sensor types based on the distance of the subject to the sensor. We expect additional differences based on the lighting and temperature conditions as the regular RGB video will lose contrast in the evening hours and turn to a black video in a night setting while the thermal sensor should continue to function well. In another setting like for example a humid green house, the atmospheric temperature might be close to the body temperature of a person and thus reduce the detection performance of the thermal sensor. As would be expected, subject orientation further has a big influence on detection performance.

These considerations show us that datasets fitting the task area, sensor setup, and recording scenario are crucial to the development of algorithms applicable in real life.

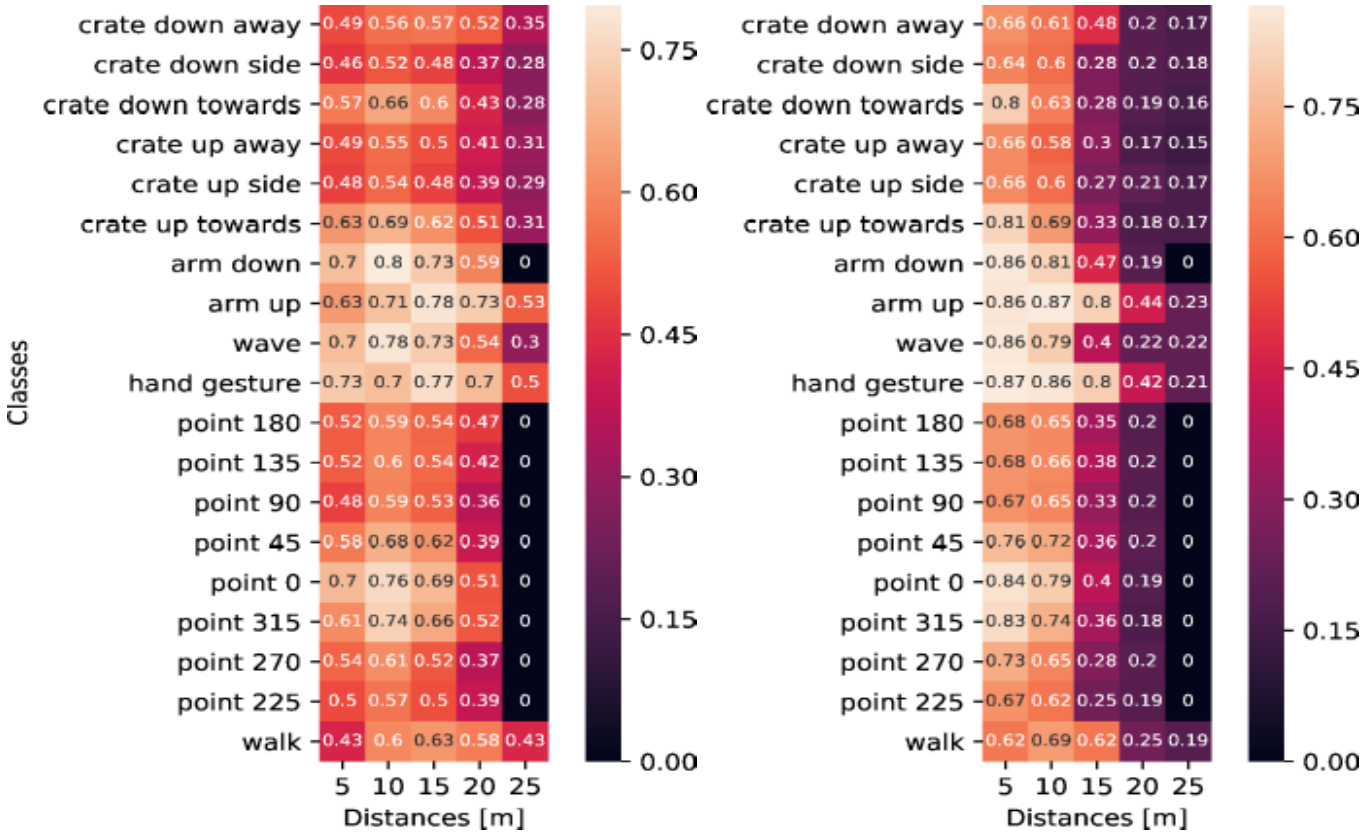

Figure 4: Average Skeleton detection confidence for Optris thermal sensor on the left and ZED RGB-D sensor (single RGB video) on the right. Distances on the X-axis from 5m to 25m, confidence values ranging from 0 to 1. Notable is the higher performance of the ZED camera at short range, but lower performance at long range. Also notable is the performance dependence on viewpoint. Actions facing the camera are generally captured better.

The recognition of actions is an important aspect of interacting with humans. However, this only encompasses the overt behavior of the humans in the vicinity of the autonomous robots. Equally important is the identification of the (covert) intentions of the humans when acting. It is from these that the robots would be best able to plan an appropriate response, whether this is providing physical assistance (e.g. moving to the appropriate location) or enhancing safety (e.g. proactively moving out of the way). Our goal in establishing the data processing pipeline, whose beginning is introduced in this paper, is to provide the data to address the issue of intention recognition. We will proceed to integrate more sensors which should lead to more robust pose estimation over a greater variety of conditions. In the case of the 3D LIDAR, we expect to gain approximate pose estimation for subjects outside the field of view of the directional sensors. The pipeline will further be supplemented with contextual information drawn from other robot systems, such as navigation, mapping, scheduling, etc.

The completed system should react to commands given by workers, track individual worker progress towards a full crate to preemptively navigate toward the next task, and learn individual worker's preferences when it comes to a comfortable stopping distance.

ACKNOWLEDGMENT

We thank the RASberry project (https://rasberryproject.com) [5] and its researchers as well as our voluntary actors for their support and cooperation.

REFERENCES

[1] W. Kay, K. Simonyan, B. Zhang, C. Hillier, F. Viola, T. Green, T. Back, P. Natsev, and A. Zisserman, "The Kinetics Human Action Video Dataset.", 2017.

[2] A. Shahroudy, J. Liu, T.-t. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," 2016.

[3] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding." in CVPR, pp. 961–970, 2015.

[4] Z. Pezzementi, et. al., "Comparing apples and oranges: Off-road pedestrian detection on the National Robotics Engineering Center agricultural person-detection dataset," Journal of Field Robotics, vol. 35, no. 4, pp. 545–563, 2018.

[5] P. From, L. Grimstad, M. Hanheide, S. Pearson, and G. Cielniak, "RASberry - Robotic and Autonomous Systems for Berry Production." Mechanical Engineering Magazine Select Articles, vol. *140, no.* 6, pp. 14-18, 2018.

[6] L. Grimstadt, and P. From, "The Thorvald II agricultural robotic system." Robotics, vol. 6, no. 4, p. 24, 2017.

[7] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image," CVPR, pp. 5689–5698, 2017.

[8] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," CVPR, pp. 4724–4732, 2016.

[9] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," CVPR, pp. 1302–1310, 2017.

[10] X. Gu, F. Deligianni, B. Lo, W. Chen, and G. Yang, "Markerless Gait Analysis Based on a Single RGB Camera," IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks, pp. 42–45, 2018.

[11] J. Wilms, G. Beckers, T. Callemein, L. Geurts, and T. Goedemé, "Human Pose Matching", Dortmund International Research Conference, pp. 65–69, 2018.